Adversarial Risk Bounds through Sparsity based Compression (Supplemetary Material)

Emilio Rafael Balda
RWTH Aachen University

Niklas Koep
RWTH Aachen University

Arash Behboodi
RWTH Aachen University

Rudolf Mathar
RWTH Aachen University

A Deferred Proofs

Proof (Theorem 2.2). Since $\widehat{L}_0^\varepsilon(g_A)$ is an average of m i.i.d random variables with expectation equal to $L_0^\varepsilon(g_A)$ we may use Hoeffdingen's inequality, yielding

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\widehat{L}_0^\varepsilon(g_A) - L_0^\varepsilon(g_A) \geq \tau\right] \leq \exp\left(-2m\tau^2\right) \,.$$

Note that $|\mathcal{A}| = \exp(\log|\mathcal{A}|)$. Then, let us choose $\tau = \sqrt{\frac{\log|\mathcal{A}|}{m}}$ and take an union bound over all $A \in \mathcal{A}$, leading to

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\widehat{L}_0^\varepsilon(g_A) - L_0^\varepsilon(g_A) \geq \sqrt{\frac{\log|\mathcal{A}|}{m}}\right] \\ \leq \exp(-\log|\mathcal{A}|) \,.$$

Since f is $(\gamma, \varepsilon, \mathcal{S})$-compressible via g, then

$$\forall \mathbf{x} \in \mathcal{S} : \quad |\ell_\varepsilon(f;\mathbf{x},y) - \ell_\varepsilon(g_A;\mathbf{x},y)| \leq \gamma \,,$$

which implies that

$$\widehat{L}_0^\varepsilon(g_A) \leq \widehat{L}_\gamma^\varepsilon(f) \,.$$

Combining these results we get that

$$L_0^\varepsilon(g_A) \leq \widehat{L}_\gamma^\varepsilon(f) + \mathcal{O}\left(\sqrt{\frac{\log|\mathcal{A}|}{m}}\right)$$

with probability at least $1-\exp(-\log|\mathcal{A}|) = 1 - 1/|\mathcal{A}|$, which we consider as high probability. □

Proof (Lemma 3.2). Note that $\mathbb{E}[\widehat{w}_i] = \frac{w_i}{p_i}\mathbb{E}[z_i] = w_i$ thus $\mathbb{E}[\widehat{\mathbf{w}}] = \mathbf{w}$. Similarly, $\mathbb{E}[|\widehat{w}_i|] = \left|\frac{w_i}{p_i}\right|\mathbb{E}[z_i] = |w_i|$ and since $\widehat{w}_i$'s are independent, we get $\mathbb{E}[\|\widehat{\mathbf{w}}\|_1] = \|\mathbf{w}\|_1$. This implies that

$$\begin{aligned}\mathbb{E}\left[\ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right] &= \mathbb{E}\left[\langle\widehat{\mathbf{w}},\mathbf{x}\rangle - \varepsilon\|\widehat{\mathbf{w}}\|_1\right] \\ &= \langle\mathbf{w},\mathbf{x}\rangle - \varepsilon\|\mathbf{w}\|_1 = \ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y)\,.\end{aligned}$$

Now lets compute the variance of $\widehat{w}_i$ as

$$\begin{aligned}\mathrm{Var}\left[\widehat{w}_i\right] &= \mathbb{E}\left[\widehat{w}_i^2\right] - \mathbb{E}\left[\widehat{w}_i\right]^2 \\ &= (w_i/p_i)^2 p_i - w_i^2 = \frac{1-p_i}{p_i}w_i^2 \,.\end{aligned}$$

The same calculation yields

$$\mathrm{Var}\left[|\widehat{w}_i|\right] = \frac{1-p_i}{p_i}w_i^2 \,.$$

The covariance between $|\widehat{w}_i|$ and $\widehat{w}_i$ is

$$\begin{aligned}\mathrm{Cov}\left(|\widehat{w}_i|, \widehat{w}_i\right) &= \mathbb{E}\left[|\widehat{w}_i|\widehat{w}_i\right] - \mathbb{E}[|\widehat{w}_i|]\mathbb{E}[\widehat{w}_i] \\ &= \frac{1-p_i}{p_i}|w_i| \cdot w_i \,.\end{aligned}$$

Now putting all together we get

$$\begin{aligned}&\mathrm{Var}\left[\widehat{w}_i x_i - \varepsilon|\widehat{w}_i|\right] \\ &\quad = x_i^2\mathrm{Var}[\widehat{w}_i] - 2\varepsilon x_i\mathrm{Cov}(\widehat{w}_i, |\widehat{w}_i|) + \varepsilon^2\mathrm{Var}[|w_i|^2] \\ &\quad = \frac{1-p_i}{p_i}\left(x_i^2 w_i^2 - 2\varepsilon x_i|w_i|w_i + \varepsilon^2 w_i^2\right) \\ &\quad \leq \frac{w_i^2}{p_i}\left(x_i^2 + 2\varepsilon|x_i| + \varepsilon^2\right) \\ &\quad = \frac{\delta\gamma^2}{(1+\varepsilon)^2}|w_i|\left(x_i^2 + 2\varepsilon|x_i| + \varepsilon^2\right) \,.\end{aligned}$$

Since $\widehat{w}_i$'s are independent, we get

$$\begin{aligned}&\mathrm{Var}\left[\langle\widehat{\mathbf{w}},\mathbf{x}\rangle - \varepsilon\|\widehat{\mathbf{w}}\|_1\right] \\ &= \mathrm{Var}\left[\sum_{i=1}^n \widehat{w}_i x_i - \varepsilon|\widehat{w}_i|\right] \\ &= \sum_{i=1}^n \mathrm{Var}\left[\widehat{w}_i x_i - \varepsilon|\widehat{w}_i|\right] \\ &\leq \frac{\delta\gamma^2}{(1+\varepsilon)^2}\sum_{i=1}^n |w_i|\left(x_i^2 + 2\varepsilon|x_i| + \varepsilon^2\right) \\ &= \frac{\delta\gamma^2}{(1+\varepsilon)^2}\left(\left\langle|\mathbf{w}|, \mathbf{x}^2\right\rangle + 2\varepsilon\left\langle|u|, |c|\right\rangle + \varepsilon^2\|\mathbf{w}\|_1\right) \\ &\leq \frac{\delta\gamma^2}{(1+\varepsilon)^2}\left(\|\mathbf{w}\|_1\left\|\mathbf{x}^2\right\|_\infty + 2\varepsilon\|\mathbf{w}\|_1\|\mathbf{x}\|_\infty + \varepsilon^2\|\mathbf{w}\|_1\right) \\ &\leq \frac{\delta\gamma^2}{(1+\varepsilon)^2}(1 + 2\varepsilon + \varepsilon^2) = \delta\gamma^2 \,,\end{aligned}$$

where $\mathbf{x}^2$ denotes the entry-wise raise to the power of 2. By Chebyshev's inequality we get

$$\mathbb{P}\left[|(\langle\widehat{\mathbf{w}},\mathbf{x}\rangle - \varepsilon\|\widehat{\mathbf{w}}\|_1) - \langle\mathbf{w},\mathbf{x}\rangle - \varepsilon\|\mathbf{w}\|_1| > \gamma\right] \leq \delta \,.$$

On the other hand, the expected number of non-zero entries in $\widehat{\mathbf{w}}$ is given by

$$\mathbb{E}\left[\|\widehat{\mathbf{w}}\|_0\right] = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \frac{|w_i|}{\delta\gamma^2}(1+\varepsilon^2) = \frac{(1+\varepsilon)^2}{\delta\gamma^2}\,.$$

Then, by Hoefdingen's inequality the number of non-zero entries in $\widehat{\mathbf{w}}$ is less than $\mathcal{O}((\log n)(1+\varepsilon)^2/\delta\gamma^2)$ with high probability. □

Proof (Lemma 3.3). We start by bounding the error incurred by clipping, that is

$$\begin{aligned}
&|\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y)| \\
&\le |\langle \mathbf{w},\mathbf{x}\rangle - \langle \mathbf{w}',\mathbf{x}\rangle| + \varepsilon\,\left|\|\mathbf{w}\|_1 - \|\mathbf{w}'\|_1\right| \\
&\le |\langle \mathbf{w}-\mathbf{w}',\mathbf{x}\rangle| + \varepsilon\,\|\mathbf{w}-\mathbf{w}'\|_1 \\
&\le \|\mathbf{w}-\mathbf{w}'\|_1\,\|\mathbf{x}\|_\infty + \varepsilon\,\|\mathbf{w}-\mathbf{w}'\|_1 \\
&\le \|\mathbf{w}-\mathbf{w}'\|_1\,(1+\varepsilon) \\
&\le \frac{\gamma}{4n(1+\varepsilon)} n(1+\varepsilon) = \gamma/4\,.
\end{aligned}$$

Similarly, the error incurred by discretizing $\widetilde{\mathbf{w}}$ is bounded by

$$\begin{aligned}
\left|\ell_\varepsilon(f_{\widetilde{\mathbf{w}}};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right| &\le \|\widetilde{\mathbf{w}}-\widehat{\mathbf{w}}\|_1\,(1+\varepsilon) \\
&\le \frac{\gamma}{2n(1+\varepsilon)}\frac{n}{2}(1+\varepsilon) \\
&= \gamma/4\,.
\end{aligned}$$

By Lemma 3.2, we know that with probability at least $1-\delta$ we have that

$$\left|\ell_\varepsilon(f_{\widetilde{\mathbf{w}}};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right| \le \gamma/2\,.$$

Combining these three results yields

$$\begin{aligned}
&|\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)| \\
&\qquad\le \quad |\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y)| \\
&\qquad\quad + \left|\ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y) - \ell_\varepsilon(f_{\widetilde{\mathbf{w}}};\mathbf{x},y)\right| \\
&\qquad\quad + \left|\ell_\varepsilon(f_{\widetilde{\mathbf{w}}};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right| \\
&\qquad\le \quad \gamma/4 + \gamma/2 + \gamma/4 \le \gamma
\end{aligned}$$

with probability at least $1-\delta$. □

Proof (Theorem 3.4). Let $\mathcal{A}$ be the set of vectors with at most $\mathcal{O}((\log n)(1+\varepsilon)^2/\delta\gamma^2)$ non-zero entries, where each entry is a multiple of $2\gamma/2n(1+\varepsilon)$ between $-\delta\gamma^2/(1+\varepsilon)^2$ and $\delta\gamma^2/(1+\varepsilon)^2$. Then, $|\mathcal{A}| = r^q$ with

$$r = 2\frac{\delta\gamma^2/(1+\varepsilon)^2}{2\gamma/2n(1+\varepsilon)} = \frac{4n\delta\gamma}{(1+\varepsilon)}\,, \qquad q = \frac{(1+\varepsilon)^2}{\delta\gamma^2}\,.$$

Let $\widehat{\mathbf{w}}$ be defined as in Lemma 3.3. Then, by Lemma 3.2, we have that $\mathbb{P}_{\widehat{\mathbf{w}}}[\widehat{\mathbf{w}} \in \mathcal{A}] \le 1-\delta$. We define $\mathcal{G} = \{f_{\widehat{\mathbf{w}}} : \widehat{\mathbf{w}} \in \mathcal{A}\}$. Note that the mapping from $f_{\mathbf{w}}$ to $f_{\widehat{\mathbf{w}}}$ fails (*i.e.*, $\widehat{\mathbf{w}} \notin \mathcal{A}$) with probability at most δ, thus corollary 2.2.1 yields

$$\begin{aligned}
&L_0^\varepsilon(f_{\widehat{\mathbf{w}}}) \\
&\le \widehat{L}_\gamma^\varepsilon(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{(1+\varepsilon)^2\log(n)\log\left(\frac{4n\delta\gamma}{(1+\varepsilon)}\right)}{\delta\gamma^2 m}}\right) + \delta \\
&= \widehat{L}_\gamma^\varepsilon(f_{\mathbf{w}}) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{(1+\varepsilon)^2}{\delta\gamma^2 m}}\right) + \delta
\end{aligned}$$

with high probability. Then, we choose $\delta = ((1+\varepsilon)^2/\gamma^2 m)^{1/3}$ which leads to

$$L_0^\varepsilon(f_{\widehat{\mathbf{w}}}) \le \widehat{L}_\gamma^\varepsilon(f_{\mathbf{w}}) + \widetilde{\mathcal{O}}\left(\left(\frac{(1+\varepsilon)^2}{\gamma^2 m}\right)^{1/3}\right)$$

with high probability. □

Proof (Lemma 3.7). Let us first bound how much does sparsifying $\mathbf{w}$ affects inner products, that is

$$|\langle \mathbf{w},\mathbf{x}\rangle - \langle \mathbf{w}',\mathbf{x}\rangle| \le \|\mathbf{w}-\mathbf{w}'\|_1\,\|\mathbf{x}\|_\infty \le \|\mathbf{w}-\mathbf{w}'\|_1\,.$$

This distorts the adversarial margin as follows:

$$\begin{aligned}
&|\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y)| \\
&\le |\langle \mathbf{w},\mathbf{x}\rangle - \langle \mathbf{w}',\mathbf{x}\rangle| + \varepsilon\,\left|\|\mathbf{w}\|_1 - \|\mathbf{w}'\|_1\right| \\
&\le \|\mathbf{w}-\mathbf{w}'\|_1 + \varepsilon\,\|\mathbf{w}-\mathbf{w}'\|_1 && \text{(triangle inequality)} \\
&= (1+\varepsilon)\,\|\mathbf{w}-\mathbf{w}'\|_1 \\
&\le (1+\varepsilon)\frac{1}{4s}\,\|\mathbf{w}\|_{1/2} && \text{(Lemma 3.6)} \\
&\le (1+\varepsilon)\frac{\overline{s}}{4s}\,\|\mathbf{w}\|_1 && \text{(Definition 3.5)} \\
&= \gamma/2\,. && \text{(Choice of } s\text{)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\left|\ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right| &\le (1+\varepsilon)\,\|\mathbf{w}'-\widehat{\mathbf{w}}\|_1 \\
&\le (1+\varepsilon)s\frac{1}{2}\left(\frac{\gamma}{s(1+\varepsilon)}\right) \\
&= \gamma/2\,.
\end{aligned}$$

Putting all together, we get

$$\begin{aligned}
&|\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)| \\
&\qquad\le |\ell_\varepsilon(f_{\mathbf{w}};\mathbf{x},y) - \ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y)| \\
&\qquad\quad + \left|\ell_\varepsilon(f_{\mathbf{w}'};\mathbf{x},y) - \ell_\varepsilon(f_{\widehat{\mathbf{w}}};\mathbf{x},y)\right| \\
&\qquad\le \gamma/2 + \gamma/2 \,= \gamma\,,
\end{aligned}$$

which completes the proof. □

Proof (Theorem 3.8). Let $\mathcal{A}$ be the set of vectors with at most $\overline{s}(1+\varepsilon)/2\gamma$ non-zero entries, where each entry is a multiple of $\gamma/s(1+\varepsilon)$ between -1 and 1. Then, $|\mathcal{A}| = r^q$ with

$$r = \frac{2}{\gamma/s(1+\varepsilon)} = \frac{2}{\gamma^2/2\overline{s}(1+\varepsilon)^2} = \frac{4\overline{s}(1+\varepsilon)^2}{\gamma^2}$$

and

$$q = s = \overline{s}(1+\varepsilon)/2\gamma\,.$$

Let us define the set

$$\mathcal{G} = \{f_{\widehat{\mathbf{w}}} : \widehat{\mathbf{w}} \text{ defined as in Lemma 3.7 with } \mathbf{w} \in \mathcal{B}^n_{1,1}\}.$$

Then, by Lemma 3.7 we know that $f_{\mathbf{w}}$ is $(\gamma, \varepsilon, \mathcal{S})$-compressible via $\mathcal{G}$, thus Theorem 2.2 yields

$$\begin{aligned} L_0^\varepsilon(f_{\widehat{\mathbf{w}}}) &\le \widehat{L}_\gamma^\varepsilon(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{2\overline{s}(1+\varepsilon)\log\left(\frac{4\overline{s}(1+\varepsilon)^2}{\gamma^2}\right)}{\gamma m}}\right) \\ &= \widehat{L}_\gamma^\varepsilon(f_{\mathbf{w}}) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{(1+\varepsilon)\overline{s}}{\gamma m}}\right) \end{aligned}$$

with high probability. □

The following lemma allows us to quantify how much error is incurred by perturbing the input of a layer, or by switching the matrix $\boldsymbol{W}$ to a different one.

Lemma A.1. *If ϕ is a 1-Lipschitz activation function, then for any $\boldsymbol{W}, \widehat{\boldsymbol{W}}$ the following inequalities hold*

$$\begin{aligned} \left\|\phi(\boldsymbol{W}^\top \mathbf{x}) - \phi(\boldsymbol{W}^\top (\mathbf{x} + \boldsymbol{\eta}))\right\|_\infty &\le \|\boldsymbol{W}\|_{1,\infty} \|\boldsymbol{\eta}\|_\infty\,, \\ \left\|\phi(\boldsymbol{W}^\top \mathbf{x}) - \phi(\widehat{\boldsymbol{W}}^\top \mathbf{x})\right\|_\infty &\le \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_{1,\infty} \|\mathbf{x}\|_\infty\,. \end{aligned}$$

Given this Lipschitz condition, proving this Lemma is trivial. Nevertheless, we provide the proof for completeness.

Proof (Lemma A.1). Since ϕ is 1-Lipschitz, we have that for any vector $\mathbf{w}$ of the same size as $\boldsymbol{\eta}$, it holds

$$|\phi(\langle \mathbf{w}, \mathbf{x}\rangle) - \phi(\langle \mathbf{w}, \mathbf{x} + \boldsymbol{\eta}\rangle)| \le |\langle \mathbf{w}, \boldsymbol{\eta}\rangle| \le \|\mathbf{w}\|_1 \|\boldsymbol{\eta}\|_\infty\,.$$

This proves the first inequality of the lemma. Similarly, for any $\mathbf{w}$ and $\widehat{\mathbf{w}}$ it follows

$$\begin{aligned} |\phi(\langle \mathbf{w}, \mathbf{x}\rangle) - \phi(\langle \widehat{\mathbf{w}}, \mathbf{x}\rangle)| &\le |\langle \mathbf{w} - \widehat{\mathbf{w}}, \mathbf{x}\rangle| \\ &\le \|\mathbf{w} - \widehat{\mathbf{w}}\|_1 \|\mathbf{x}\|_\infty\,, \end{aligned}$$

thus implying the second inequality. □

Proof (Lemma 3.10). Since $\boldsymbol{W}$ is effectively joint sparse, we can bound $\left\|\boldsymbol{W} - \overline{\boldsymbol{W}}\right\|_{1,\infty}$ as follows

$$\begin{aligned} \left\|\boldsymbol{W} - \overline{\boldsymbol{W}}\right\|_{1,\infty} &\le \frac{1}{s_2}\|\boldsymbol{W}\|_{1,1} && \text{(Lemma 3.6)} \\ &\le \frac{\overline{s}_2}{s_2}\|\boldsymbol{W}\|_{1,\infty}\,. && \text{(Definition 3.9)} \end{aligned}$$

Similarly, since the remaining non-zero columns $\overline{\boldsymbol{W}}$ are effectively sparse, we get

$$\begin{aligned} \left\|\overline{\boldsymbol{W}} - \widetilde{\boldsymbol{W}}\right\|_{1,\infty} &= \inf_{\boldsymbol{X}:\|\boldsymbol{X}\|_{0,\infty} = s_1} \left\|\overline{\boldsymbol{W}} - \boldsymbol{X}\right\|_{1,\infty} \\ &\le \frac{1}{4s_1}\left\|\overline{\boldsymbol{W}}\right\|_{1/2,\infty} && \text{(Lemma 3.6)} \\ &\le \frac{\overline{s}_1}{4s_1}\left\|\overline{\boldsymbol{W}}\right\|_{1,\infty}\,. && \text{(Definition 3.5)} \end{aligned}$$

By the definition of $\widehat{\boldsymbol{W}}$, we have that $\left\|\widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}}\right\|_{1,\infty} \le \gamma/3$. Combining all these statements, the choice of s_1 and s_2 (see Algorithm 1) yields

$$\begin{aligned} \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_{1,\infty} &\le \left\|\boldsymbol{W} - \overline{\boldsymbol{W}}\right\|_{1,\infty} \\ &\quad + \left\|\overline{\boldsymbol{W}} - \widetilde{\boldsymbol{W}}\right\|_{1,\infty} \\ &\quad + \left\|\widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}}\right\|_{1,\infty} \\ &\le \frac{\overline{s}_1}{4s_1}\|\boldsymbol{W}\|_{1,\infty} + \frac{\overline{s}_2}{s_2}\|\boldsymbol{W}\|_{1,\infty} + \frac{\gamma}{3} \\ &\le \frac{\gamma}{3} + \frac{\gamma}{3} + \frac{\gamma}{3} = \gamma\,. \end{aligned}$$

It remains to bound the covering number of $\mathcal{W}$ with the mixed $(1,\infty)$-norm, denoted by $\mathcal{N}(\mathcal{W}, \|\cdot\|_{1,\infty}, \gamma/3)$. By definition, the set $\mathcal{W}$ is composed of all matrices $\widetilde{\boldsymbol{W}}$ with at most s_2 non-zero columns, where each column has at most s_1 non-zero entries and ℓ_1-norm not greater than one. Since any $\widetilde{\boldsymbol{W}} \in \mathcal{W}$ has at most s_2 non-zero columns, we get

$$\begin{aligned} &\mathcal{N}(\mathcal{W}, \|\cdot\|_{1,\infty}, \gamma/3) \\ &\le \binom{n_2}{s_2}\mathcal{N}(\gamma/3, \mathcal{B}^{n_1}_{1,1} \cap \mathcal{B}^{n_1}_{0,s_1}, \|\cdot\|_1)^{s_2} \\ &\le \binom{n_2}{s_2}\left[\binom{n_1}{s_1}\mathcal{N}(\gamma/3, \mathcal{B}^{s_1}_{1,1}, \|\cdot\|_1)\right]^{s_2} \\ &\le \left(\frac{en_2}{s_2}\right)^{s_2}\left[\left(\frac{en_1}{s_1}\right)^{s_1}\mathcal{N}(\gamma/3, \mathcal{B}^{s_1}_{1,1}, \|\cdot\|_1)\right]^{s_2} \\ &\le \left(\frac{en_2}{s_2}\right)^{s_2}\left(\frac{en_1}{s_1}\right)^{s_1 s_2}\left(1 + \frac{6}{\gamma}\right)^{s_1 s_2}\,. \end{aligned}$$

This leads to

$$\begin{aligned} \log \mathcal{N}(\mathcal{W}, \|\cdot\|_{1,\infty}, \gamma/3) &\le \widetilde{\mathcal{O}}\left(s_1 s_2\right) \\ &= \widetilde{\mathcal{O}}\left(\|\boldsymbol{W}\|^2_{1,\infty}\,\overline{s}_1\overline{s}_2/\gamma^2\right)\,. \end{aligned}$$

choosing $\mathcal{C}$ to be the covering set of $\mathcal{W}$ completes the proof. □

Proof (Theorem 3.11). By assumption, the activation functions are all set to be the ReLU activation ϕ. Then, due to its positive homogeneity property, we rebalance the network by setting $\left\|\boldsymbol{W}^i\right\|_{1,\infty} = 1$ for all $i = 1, \dots, d$ without altering the classification function. For any given adversarial noise $\boldsymbol{\eta}_1$ with ℓ_∞-norm bounded by ε, let us re-define $\mathbf{x}^i$ as in (1) but with $\mathbf{x}^0 = \mathbf{x} + \boldsymbol{\eta}_1$. Similarly, for another adversarial noise $\boldsymbol{\eta}_2$ with ℓ_∞-norm bounded by ε and compressed matrices $\widehat{\boldsymbol{W}}^i$, let us define the error vector of the i-th layer $\boldsymbol{\eta}^i$ in a recursive fashion, that is $\boldsymbol{\eta}^i := \phi(\boldsymbol{W}^{i^\top}\mathbf{x}^{i-1}) - \phi(\widehat{\boldsymbol{W}}^{i^\top}(\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1}))$ for $i = 1, \dots, d$ with $\boldsymbol{\eta}^0 := \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1$. Note that $\left\|\boldsymbol{\eta}^0\right\|_\infty \le 2\varepsilon$. With this definition of $\mathbf{x}^i$, since

$$\begin{aligned}\left\|\phi(\boldsymbol{W}^{i^\top}\mathbf{x}^{i-1})\right\|_\infty &\le \left\|\boldsymbol{W}^{i^\top}\mathbf{x}^{i-1}\right\|_\infty \\ &\le \left\|\boldsymbol{W}^{i^\top}\right\|_\infty \left\|\mathbf{x}^{i-1}\right\|_\infty \\ &= \left\|\boldsymbol{W}^{i}\right\|_{1,\infty} \left\|\mathbf{x}^{i-1}\right\|_\infty\end{aligned}$$

we have that $\left\|\mathbf{x}^i\right\|_\infty \le \left\|\mathbf{x}^0\right\|_\infty \prod_{j=1}^{i} \left\|\boldsymbol{W}^j\right\|_{1,\infty} \le 1+\varepsilon$.

Our first goal is to bound $\left\|\boldsymbol{\eta}^i\right\|_\infty$ for $i = 1, 2, \dots, d$, which we do by induction. For any $i > 0$, let us assume that $\left\|\boldsymbol{\eta}^{i-1}\right\| \le \varepsilon^{i-1}$ where ε^{i-1} is some positive value. Given some $\varepsilon^i > \varepsilon^{i-1}$, we compress $\boldsymbol{W}^i$ as $\widehat{\boldsymbol{W}}^i = \mathrm{MatrixCompress}((\varepsilon^i - \varepsilon^{i-1})/(1+\varepsilon+\varepsilon^{i-1}), \boldsymbol{W}^i)$. Then, using Lemma A.1, we get

$$\begin{aligned}&\left\|\boldsymbol{\eta}^i\right\|_\infty = \left\|\phi(\boldsymbol{W}^{i^\top}\mathbf{x}^{i-1}) - \phi(\widehat{\boldsymbol{W}}^{i^\top}(\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1}))\right\|_\infty \\ &= \quad \left\|\phi(\boldsymbol{W}^{i^\top}\mathbf{x}^{i-1}) - \phi(\boldsymbol{W}^{i^\top}(\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1}))\right\|_\infty \\ &+ \left\|\phi(\boldsymbol{W}^{i^\top}(\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1})) - \phi(\widehat{\boldsymbol{W}}^{i^\top}(\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1}))\right\|_\infty \\ &\le \left\|\boldsymbol{W}^i\right\|_{1,\infty} \left\|\boldsymbol{\eta}^{i-1}\right\|_\infty \\ &\quad + \left\|\boldsymbol{W}^i - \widehat{\boldsymbol{W}}^i\right\|_{1,\infty} \left\|\mathbf{x}^{i-1} + \boldsymbol{\eta}^{i-1}\right\|_\infty \quad \text{(Lemma A.1)} \\ &\le \varepsilon^{i-1} + \left\|\boldsymbol{W}^i - \widehat{\boldsymbol{W}}^i\right\|_{1,\infty} (1 + \varepsilon + \varepsilon^{i-1}) \\ &\le \varepsilon^i \,. \qquad \text{(Definition of } \widehat{\boldsymbol{W}}^i\text{)}\end{aligned}$$

Given y and $f_{\boldsymbol{W}}$, let us define $\widetilde{f}_{\boldsymbol{W}}(\mathbf{x}) := [f_{\boldsymbol{W}}(\mathbf{x})]_{j \neq y}$. By setting $\varepsilon^0 := 2\varepsilon$ and $\varepsilon^d := \gamma/2$, we get

$$\begin{aligned}&\left|\ell_\varepsilon(f_{\boldsymbol{W}}; \mathbf{x}, y) - \ell_\varepsilon(f_{\widehat{\boldsymbol{W}}}; \mathbf{x}, y)\right| \\ &= \Big|[f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)]_y - \max_{j \neq y}[f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)]_j \\ &\quad - [f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)]_y + \max_{j \neq y}[f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)]_j\Big| \\ &= \Big|[f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)]_y - \left\|\widetilde{f}_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)\right\|_\infty \\ &\quad - [f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)]_y + \left\|\widetilde{f}_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)\right\|_\infty\Big| \\ &\le \left|[f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)]_y - [f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)]_y\right| \\ &\quad + \left|\left\|\widetilde{f}_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1)\right\|_\infty - \left\|\widetilde{f}_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)\right\|_\infty\right| \\ &\le \left\|f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1) - f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)\right\|_\infty \\ &\quad + \left\|\widetilde{f}_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1) - \widetilde{f}_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)\right\|_\infty \\ &\le 2\left\|f_{\boldsymbol{W}}(\mathbf{x} + \boldsymbol{\eta}_1) - f_{\widehat{\boldsymbol{W}}}(\mathbf{x} + \boldsymbol{\eta}_2)\right\|_\infty \\ &= 2\left\|\boldsymbol{\eta}^d\right\|_\infty \le \gamma \,.\end{aligned}$$

We are free to choose $\varepsilon^1, \dots, \varepsilon^{d-1}$ without loosing this bound on $\left|\ell_\varepsilon(f_{\boldsymbol{W}}; \mathbf{x}, y) - \ell_\varepsilon(f_{\widehat{\boldsymbol{W}}}; \mathbf{x}, y)\right|$, as long as $\varepsilon^i > \varepsilon^{i-1}$. However, the choice of these values will determine the sample complexity of the compressed function class. A naive way of choosing ε^i, like $\varepsilon^i := i(\gamma/2 - 2\varepsilon)/d + 2\varepsilon$, will lead sample complexity of $\mathcal{O}(d^2)$ instead of $\mathcal{O}(d)$. Therefore, we choose these parameters in a smarter way, that is

$$\varepsilon^0 := 2\varepsilon \,, \qquad \varepsilon^i := \varepsilon^{i-1} + \frac{\sqrt{\overline{s}_1^i \overline{s}_2^i}}{\sum_{j=1}^{d} \sqrt{\overline{s}_1^j \overline{s}_2^j}} (\gamma/2 - 2\varepsilon) \,,$$

so that more error is allocated to the layers with more effective parameters. Note that this selection implies $\varepsilon^d = \gamma/2$ and $\varepsilon^i > \varepsilon^{i-1}$, so $f_{\boldsymbol{W}}$ is $(\gamma, \varepsilon, \mathcal{S})$-compressible via $\mathcal{G} = \{f_{\widehat{\boldsymbol{W}}} : \widehat{\boldsymbol{W}} = \mathrm{MatrixCompress}((\varepsilon^i - \varepsilon^{i-1})/(1 + \varepsilon + \varepsilon^{i-1}), \boldsymbol{W})\}$. In the same manner as in Lemma A.1, for all $i = 1, \dots, d$ let us define $\mathcal{C}^i$ to be the set of all possible $\widehat{\boldsymbol{W}}^i$. With this choice, the logarithm of the cardinality of the compressed function class is

$$\begin{aligned}
\log|\mathcal{G}| &= \log\prod_{i=1}^{d}\left|\mathcal{C}^i\right| = \sum_{i=1}^{d}\log\left|\mathcal{C}^i\right| \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{d}\overline{s}_1^i\overline{s}_2^i(1+\varepsilon+\varepsilon^{i-1})^2/(\varepsilon^i-\varepsilon^{i-1})^2\right) \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{d}\frac{\overline{s}_1^i\overline{s}_2^i(1+\varepsilon+\gamma/2-2\varepsilon)^2\left(\sum_{j=1}^{d}\sqrt{\overline{s}_1^j\overline{s}_2^j}\right)^2}{\left((\gamma/2-2\varepsilon)\sqrt{\overline{s}_1^i\overline{s}_2^i}\right)^2}\right) \\
&= \widetilde{\mathcal{O}}\left(\sum_{i=1}^{d}\frac{(1+\gamma/2-\varepsilon)^2\left(\sum_{j=1}^{d}\sqrt{\overline{s}_1^j\overline{s}_2^j}\right)^2}{(\gamma/2-2\varepsilon)^2}\right) \\
&= \widetilde{\mathcal{O}}\left(d\left(\frac{1+\gamma/2-\varepsilon}{\gamma/2-2\varepsilon}\right)^2\left(\sum_{j=1}^{d}\sqrt{\overline{s}_1^j\overline{s}_2^j}\right)^2\right).
\end{aligned}$$

Finally, we apply Theorem 2.2, yielding

$$\begin{aligned}
&L_0^{\varepsilon}(f_{\widehat{\boldsymbol{W}}}) \leq \widehat{L}_{\gamma}^{\varepsilon}(f_{\boldsymbol{W}}) \\
&\quad + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}\left(\frac{1+\gamma/2-\varepsilon}{\gamma/2-2\varepsilon}\right)^2\left(\sum_{j=1}^{d}\sqrt{\overline{s}_1^j\overline{s}_2^j}\right)^2}\right)
\end{aligned}$$

which proves the theorem. □

B Details about Experiments

We train a fully connected neural network of 3 layers with ReLU activations on the MNIST and CIFAR-10 datasets. After preprocessing, the inputs are 1024-dimensional vectors with ℓ_∞-norm bounded by one. The weight matrices are of size 1024×500, 500×150, and 150×10. To estimate the adversarial risk, we use the projected gradient descent (PGD) attack (Madry et al. 2018) with ℓ_∞-norm bounded by 0.2 and perturbations computed through 10 iterations of the PGD algorithm. This PGD method is the state of the art algorithm for adversarial training.

In Figure 1(a), the network is first trained, on the MNIST dataset, without using adversarial examples. Then, after 50% of the training time, we start introducing adversarial examples to the training set. The same procedure is done in Figure 1(b) for the CIFAR-10 dataset, but adversarial examples are introduced after 33% of the training time. We split training into these two phases to distinguish between bounds that correlate with adversarial error and ones that correlate with standard error. These experiments are carried out using the PGD method as described above, except for 0.1 bound on the perturbation's ℓ_∞-norm. Instead, we start with a 0.05 norm bound and slowly increase it until reaching 0.1.

The source code for these experiments is available at `github.com/ebalda/adversarial-risk-bounds`